jh130016–DA01

超並列計算による経済・社会現象のビッグデータ解析

大西 立顕 （東京大学）

概要　様々なビッグデータに基づいて複雑な社会経済システムを分析した．首都圏の住宅取引の売買データを用いて，局所的な不動産バブルの度合いを定量化する手法を確立し，アベノミクス効果を観測した．国内の金融機関の振込情報を用いて，企業間の振込ネットワークを構築し，スケールフリー性，負の次数相関，PageRank と次数と振込総額の間の関係性を明らかにした．世界のビジネスニュースのテキスト時系列を分析し，ニュースの記事数が地域の活動度と相関することを明らかにした．k 近傍法を用いて商業統計メッシュデータを解析し，店舗の販売額の生産関数を売り場面積，従業員数，地価の関数として数値的に算出した．

1．研究の目的と意義

情報通信技術が飛躍的に向上したことにより，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ようになってきている．また，コンピューターの計算性能の向上に伴ない，今まではほとんど不可能に近かった，大量のデータを用いて大量の計算を行うような分析が可能になってきている．こうした背景から，特に経済物理学の分野において，経済学者だけでなく物理学や数理工学の研究者が大規模な経済データを用いて経済現象を研究するようになっている．最近では，こうした様々なビッグデータを利活用することで，経済現象のみならず社会現象を含むあらゆる現象を実証科学的に研究・分析し，社会やビジネスに役立てる試みも活発になってきている．

このような現実のビッグデータに基づいた解析を行い，経済・社会現象のモデル化・予測・制御に役立てるための新たな方法論や理論体系の構築に向けた研究を行う．経済・社会現象の科学的な理解を深め，新たな社会システムやサービスを創成し，社会の多様な課題を解決していくことを目指す．本研究では，不動産バブル，企業間の振込ネットワーク，ニュース時系列，商業統計メッシュデータの四つのテーマについて解析を行う．

2．当拠点公募型共同研究として実施した意義

(1) 共同研究を実施した拠点名および役割分担

拠点：東京大学 情報基盤センター

役割分担：

- 大西立顕 (東京大学大学院情報理工学系研究科)：代表者．研究全般の遂行．
- 渡辺努 (東京大学大学院経済学研究科)：副代表者．マクロ経済学・金融政策の視点からの支援．経済・社会データの調達．
- 清水千弘 (麗澤大学 経済学部)：不動産経済学・計量経済学の視点からの支援．空間データの調達．
- 水野貴之 (情報・システム研究機構国立情報学研究所)：経済物理学の視点からの支援．経済・社会データの調達．
- 藤本祥二 (金沢学院大学経営情報学部)：

経済物理学の視点からの支援.

- 久野遼平 (情報・システム研究機構国立情報学研究所)：統計科学の視点からの支援.

(2) 共同研究分野

超大規模データ処理系応用分野

(3) 当公募型共同研究ならではという事項など

東京大学情報基盤センター FX10 スーパーコンピュータシステムを使用できることにより，大規模計算を非常に効率的に行うことができた.

3. 研究成果の詳細と当初計画の達成状況

(1) 研究成果の詳細について

■不動産バブル

昨年，我々は首都圏の住宅取引データ (リクルート社提供．1986〜2009 年の約 73 万件) を用いて不動産バブルを分析し，物件の価格分布の地域間格差からバブルを定義する手法を提案した．住宅価格 P は，専有面積 S が広くなるほど指数的に高くなる．この関係を用いて，面積の違いを考慮した住宅価格として面積調整価格 $Q = \log P - S$ を定義した．バブルでない平常時には面積調整価格は正規分布に従うが，バブル期には価格の地域間格差が高まるため，面積調整価格は正規分布から乖離し裾野の長い分布になる．この性質を用いて，コルモゴロフ・スミルノフ検定により分布の検定を行い，面積調整価格が正規分布から乖離する度合いから同一需給圏 (価格が同質とみなせる裁定が成立する地域の広さ) の大きさを定義した．これにより，1980 年代後半に都心で発生した不動産バブルが空間的に波及し，収束していく様子を定量的に観測することができている.

アベノミクスにより，現在，日本ではデフレ

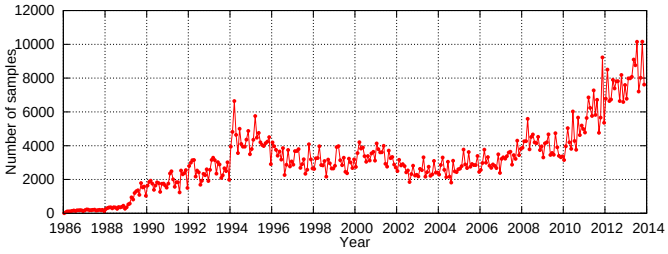

図 1: サンプル数の月次推移

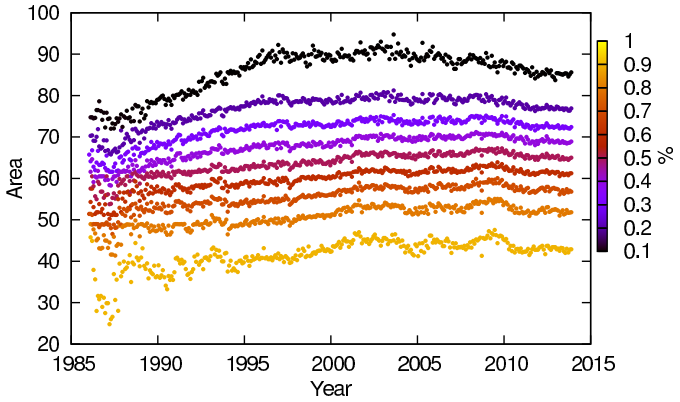

図 2: 面積の十分位数の月次推移

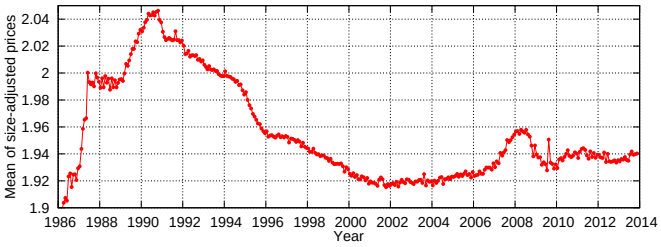

図 3: 面積調整価格の月次推移

脱却に向けて世界的にも例をみない大規模な金融緩和が行われており，日銀は 2 %の物価目標を打ち出している．この政策の成否は理論的にも実証的にも意見が分かれている．本当に安定したインフレが実現されつつあるのか？資産バブルが起きてしまっていないか？逆に地価や株価が下落し財政破綻が懸念されないか？これらのことをリアルタイムでモニターし，現状を精確に把握してリスクを早期に発見することが政策を行っていく上で重要になる．そこで，リクルート社に協力いただき，首都圏の住宅取引について最新のデータ (1986 年〜2013 年 11 月の約 100 万件) を提供していただき，アベノミクス前後で不動産バブルの度合いにどのような変化が見られるかに

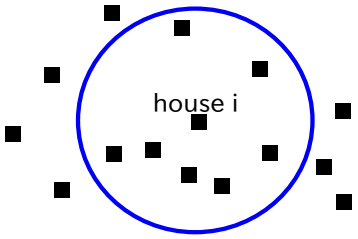

図 4: バブルの度合いの計測．青い円の中の物件が物件 i の近接物件になる $(n = 8)$.

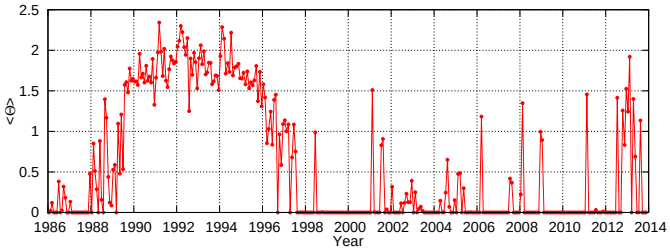

図 5: バブル度 $\langle\Theta\rangle$ の月次推移

注目して分析した．IT 技術の進展により，近年のデータは網羅性が高くなっている (図 1)．そのため，面積，価格，緯度，経度，築年数のそれぞれについて，十分位数の月次推移を観測し，サンプルバイアスがないことを確認した (図 2)．面積調整価格の月次推移は図 3 のようになっている．

ある物件 i の地点のバブルの度合い Θ_i を測るには，まず，物件 i の近接物件を $n-1$ 個を取り出し (図 4)，これら n 個の物件の面積調整価格が正規分布するかどうかを検定する．計算量を節約するため，検定にはギアリー検定 (5% 有意) を用いた．そして，物件 i の地点でのバブルの度合いを

$$\Theta_i = \log\left(\frac{\text{その時期の全物件数}}{\text{正規分布とみなせる最大の } n}\right)$$

により定義する．Θ_i は平常時は 0 になるが，バブル時には大きな値をとる．月毎に算出した結果，Θ_i をその時期の全物件で平均した $\langle\Theta\rangle$ の月次推移は図 5 のようになる．不動産バブルと言われる 1980 年台後半から 1990 年台前半にかけて $\langle\Theta\rangle$ は大きな値をとり，その後，2012 年までほぼ 0 になっている．そして，2012 年

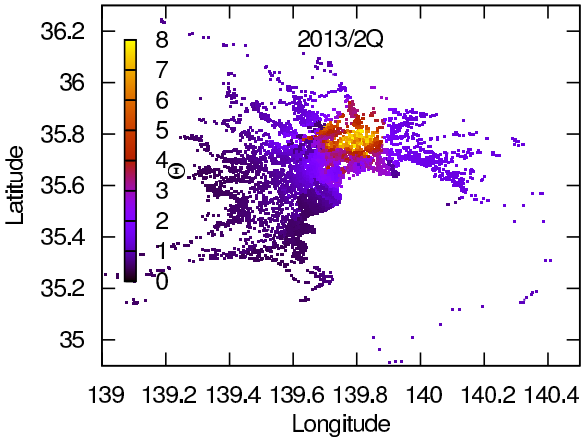

図 6: 2013 年第二四半期 (4 月〜6 月) のバブルの度合いの空間分布

後半から $\langle\Theta\rangle$ は再び継続的に大きな値をとっており，面積調整価格の分布に歪みが生じ，平常時と異なる状況になっている．特に，東京都北東部付近で Θ の大きい領域が発生していることが分かった (図 6)．実際の現場では，売りに出された物件がすぐに売れるような状況になっており，現場の感覚と矛盾しない結果になっている．ただし，2013 年第四四半期になるとこの傾向は弱まり，バブルの度合いは 0 に近づく．今後も状況をモニターし，状況を観測することが重要である．

■企業間の振込ネットワーク

国内の 3 つの金融機関に協力いただき，2012 年の一年間についての企業間の振込のデータを提供いただいた (それぞれ，データ A，データ B，データ C とする)．データには，どの企業からどの企業にいくらの金額が振込されたかについての情報が入っている．この振込データから現実のお金の流れのネットワークを構築し，ネットワーク解析を行った．企業をノードとし，企業 i から企業 j に振込がされていればノード i からノード j に有向リンクをつなぐ $(i \to j)$ という操作により，有向ネットワークを構築した．リンクの向きはお金の流れの向きになる．一般に，ネットワークはいくつかの成分に分類することができる．

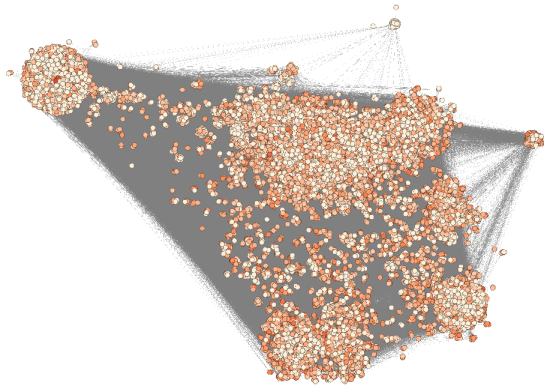

図 7: ネットワークの最大強連結成の可視化 (データ A)

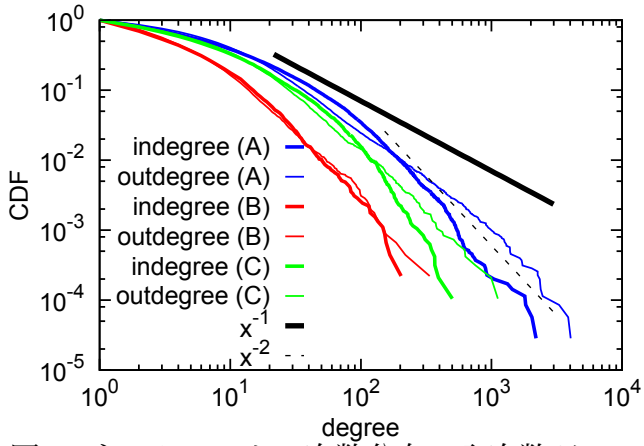

図 8: ネットワークの次数分布．入次数が in-degree，出次数が outdegree.

ここでは，強連結成分 (任意の 2 つのノード間に有向路が存在する成分) で最大のもの (最大強連結成分) のみを分析対象とする．データ A，B，C の最大強連結成分のサイズはそれぞれ約 4 万ノード，約 5 千ノード，約 1 万ノードになる (図 7).

各ノードの入次数 k^{in}(他のノードから入ってくるリンクの本数)，出次数 k^{out}(他のノードへのリンクの本数) はどちらもベキ分布に従い (図 8)，スケールフリーネットワークになっている．ただし，入次数と出次数でベキ指数は異なり，非対称なネットワークになっている．

次数 k のノードに隣接しているノードの平均次数は図 9 のような k 依存性がある．次数の大きなノードにつながっているノードの平

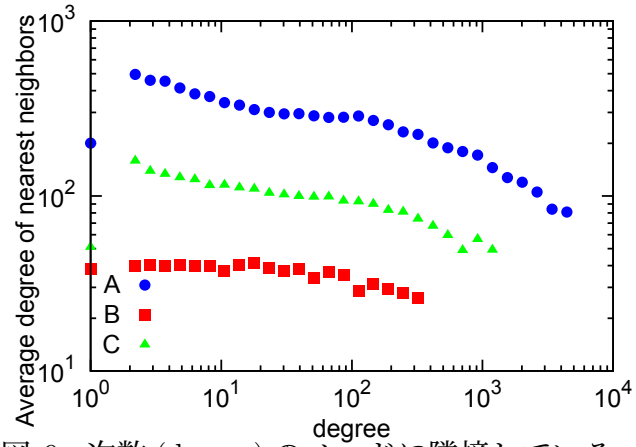

図 9: 次数 (degree) のノードに隣接しているノードの平均次数

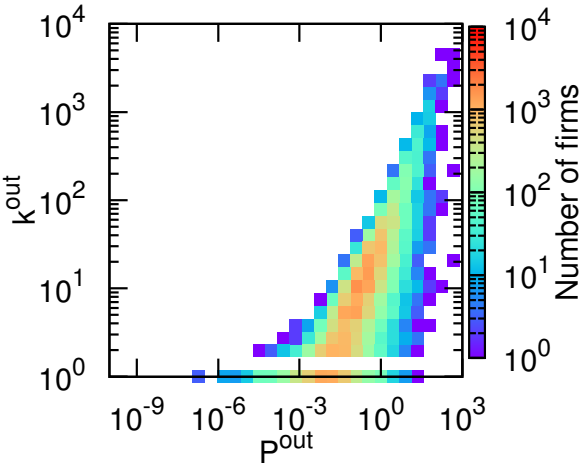

図 10: ページランク p^{out} と次数 k^{out} の散布図 (データ A)

均次数は小さく，ハブ企業同士がつながる確率は小さい．つまり，ノード間はまったくランダムにつながっているわけではなく，つながり方には相関構造がある．

ネットワークは隣接行列によって表現できる．隣接行列 A_{ij} は，ノード i からノード j へのリンクがあれば $A_{ij} = 1$，リンクがなければ $A_{ij} = 0$ と定義される (ここでは，リンクの重みの違いを考慮しない)．いま，ネットワーク上のノードをウォーカーがランダムウォークする状況を考える．ウォーカーは今いるノードのリンク先のノードのいずれかに等確率に動くとする．十分時間がたったときノード i を訪れる確率 p_i^{in} は，行列 A_{ji}/k_i^{out} の最大固有ベ

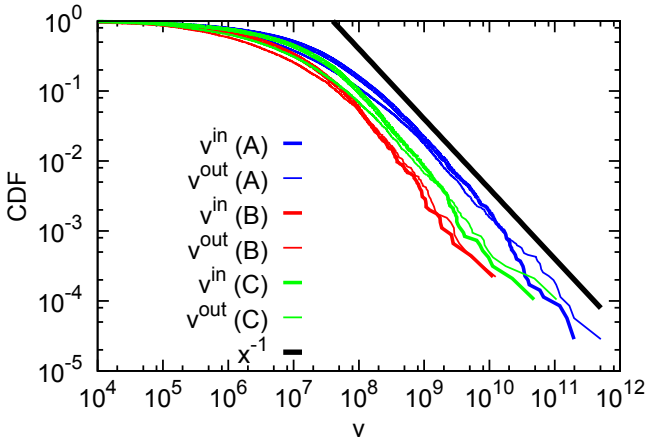

図 11: 支払った振込総額 v^{out}，受け取った振込総額 v^{in} の累積分布

クトルになる．この行列は確率行列のため最大固有値は 1 であるから，p_i^{in} は

$$p_i^{in} = \sum_j \frac{A_{ji}}{k_j^{\mathrm{out}}} p_j^{in}$$

より求まる．いま，最大強連結成分のみを解析対象としているため，p_i^{in} はノード i のページランクになる．p_i^{in} が大きな値をとるためには，ノード i の入次数が大きいだけでなく，ノード i に隣接しているノードのページランクが大きいことも重要になる．いま，ネットワークのリンクの方向を逆向きにして同様にページランク p^{out} も定義することができる．p_i^{out} は，ウォーカーがリンクを逆向きにランダムウォークするとき，ウォーカーがノード i を訪れる確率になる．これらのページランクは，ランダムウォーカーがどのくらいその企業に滞在するかを表現し，ネットワーク構造からみた企業の重要性を表す一つの指標になっている．振込ネットワークについて，ページランク $p^{\mathrm{in}}, p^{\mathrm{out}}$ を計算した．図 10 の散布図から分かるように，$p^{\mathrm{in}}, p^{\mathrm{out}}$ はそれぞれ入次数，出次数と強く相関している．Kendall の順位相関係数はそれぞれ $0.65, 0.66$ になる．ネットワークが完全にランダムにつながっていれば，ページランクと次数は同じ値になる．しかし，実

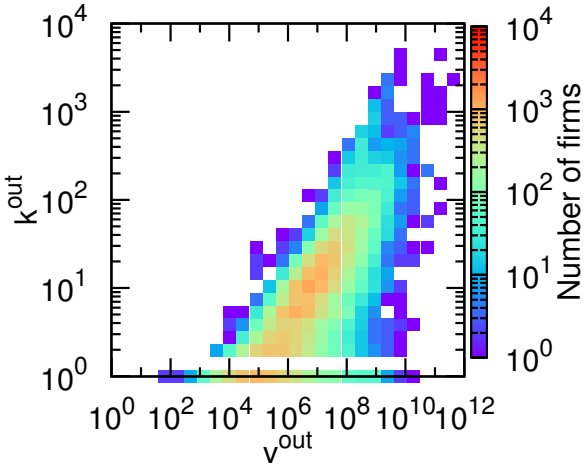

図 12: 支払った振込総額 v^{out} と次数 k^{out} の散布図 (データ A)

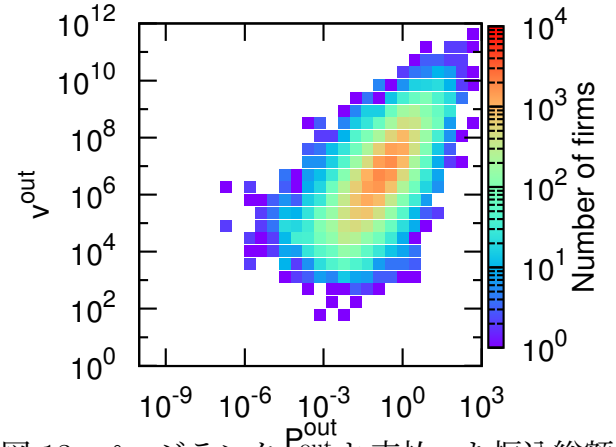

図 13: ページランク p^{out} と支払った振込総額 v^{out} の散布図 (データ A)

際にはつながり方に相関構造があるため，同じ大きさの次数を持つ企業でも，ページランクの値が異なる企業も存在している．

次に，企業間の取引金額 (振込総額) に注目し，振込総額とネットワーク構造の関係を調べた．各企業について，他の企業に支払った振込総額 v^{out}，他の企業から受け取った振込総額 v^{in} はどちらもベキ分布に従い (図 11)，各企業に出入りする金額には大きな格差がある．図 12 の散布図から分かるように，振込総額 $v^{\mathrm{in}}, v^{\mathrm{out}}$ はそれぞれ入次数 k^{in}，出次数 k^{out} と強く相関している．Kendall の順位相関係数はそれぞれ $0.43, 0.61$ になる．また，図 13 の散布図から分かるように，振込総額 $v^{\mathrm{in}}, v^{\mathrm{out}}$ はそれぞれページランク $p^{\mathrm{in}}, p^{\mathrm{out}}$ とも強く相関

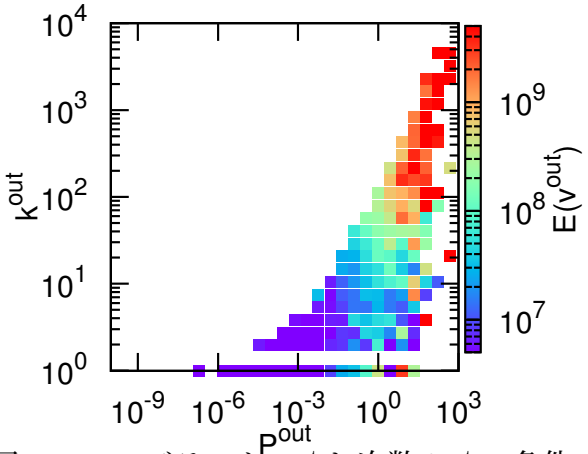

図 14: ページランク p^{out} と次数 k^{out} で条件つけした支払った振込総額 v^{out} の平均値 (データ A)

している．Kendall の順位相関係数はそれぞれ 0.35, 0.53 になる．したがって，振込総額が大きい企業は，ページランクも次数も大きい企業であることが多い．

ページランクと次数は同じような量であるが，まったく同じものではない．そこで，振込総額が，ページランクと次数の二つの情報とどう関係しているかを詳しくみるために，ページランクと次数で条件つけした振込総額の平均値 (ページランクと次数で二つの軸をとった二次元平面をセルで分割し，各セルに位置する企業についての振込総額の平均値) を調べた (図 14)．

ページランクや次数が大きい企業ほど振込総額が大きいという性質が確認できる．さらに，同じ次数の企業同士で比較をすると，ページランクの大きい企業ほど振込総額が大きいことが分かる．また，ページランクが同じ値の企業同士では，次数の大きい企業ほど振込総額が大きい．したがって，次数とページランクの両方の情報を利用することで，振込総額をより精度高く推定できる可能性がある．現実に利用できる顧客間取引データでは，顧客間の本当の取引金額が分からない場合が多い．ここで得られた知見を用いて取引金額を推定

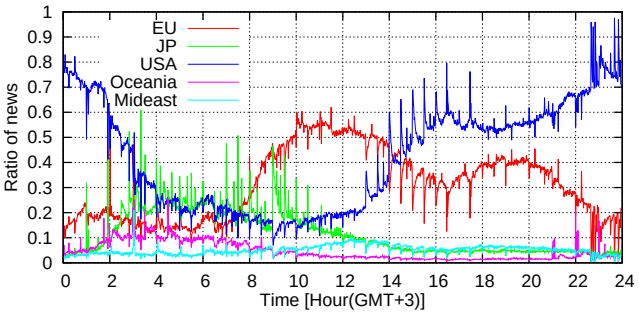

図 15: 1 分間のニュース数の割合の日中推移

できるようになれば，企業間のつながりの影響をより現実的に解析し，ネットワークの脆弱性の視点からリスクを評価することが可能になると考えられる．

■ニュース時系列

これまでの金融市場の研究では，多くの場合，価格や取引量などの数値データの時系列を対象として分析されてきた．しかし，実際の世の中は非定常なため，金融市場の分析では経済状況や外部環境の影響を考慮することも重要である．そこで，Reuters 3000 Xtra のニュースのテキストデータを用いて (English news reports のみ使用)，ニュース時系列を解析した．データは，2003〜2011 年の期間で約 2 億件の記事になる．各記事には，記事の内容を特徴づける TOPICS が振られている．これらの TOPICS のうち，地域を特徴づけるものを取り出し，EU，JP，USA，Oceania，Mideast に分類した．外国為替市場には，市場参加者の特徴を反映した日中変動のパターン (24 時間の周期性) があることが知られている．そこで，1 分毎に各地域のニュースが出現した記事数を数えた．24 時間の中の何時何分に記事が出現したかに注目し，記事数の日中推移を調べた．24 時間中の特定の時刻について，全記事数に占める各地域の記事数の割合は図 15 のようになる．日本 (JP) の記事数は，日本の活動時間帯に対応する 3〜10 時 (日本時間 9〜16 時に対応) に多くなっている．その後，EU，

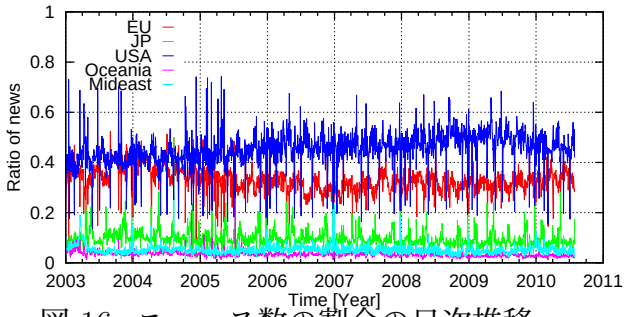

図 16: ニュース数の割合の日次推移

米国 (USA) に活動時間帯が移っていくが，記事数もそれに対応した変動をしている．したがって，記事数が地域の活動度と相関していることが確認できる．

次に，今度は日中変動を無視し，1 日毎に各地域のニュースが出現した記事数を数える．各地域の記事数の割合の日次推移は図 16 のようになる．記事数は米国が一番多く，次いで EU，日本の順になる．1 日毎にみるとどの地域も記事数は激しく変動しており，ある特定の地域の記事が普段より多くなったり少なくなったりする日が頻繁に存在している．したがって，ニュースの記事数を数えることで，ニュースの話題の中心がどこの地域にあるかを観測することが可能になる．

■商業統計メッシュデータ

2002 年の商業統計メッシュデータを用いて，全国 1,107,177 店舗について各店舗の販売額 S，売り場面積 A，従業員数 E，緯度・経度のデータを分析した．商業地に関する公示地価価格のデータを用いて，店舗の位置 (緯度・経度) に一番近い公示地価の基準地を抽出し，店舗の地価 (平米単価)L を算出した．各量の分布はベキ分布に従っている (図 17)．各量の相関関係は図 18 のようになる．S と E，S と A は強く相関し，S と L との相関は弱い．また，E と A は強く相関している．

店舗の生産性を考えるために，S がどのように決まるか，つまり，生産関数 $S = S(A, E, L)$

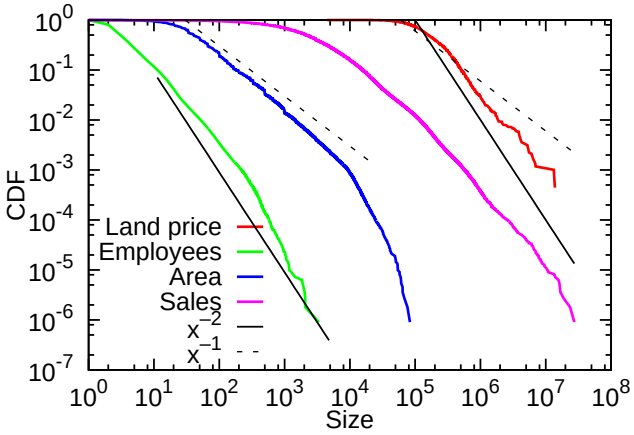

図 17: 店舗の地価 (平米単価)，従業員数，売り場面積，販売額の累積分布

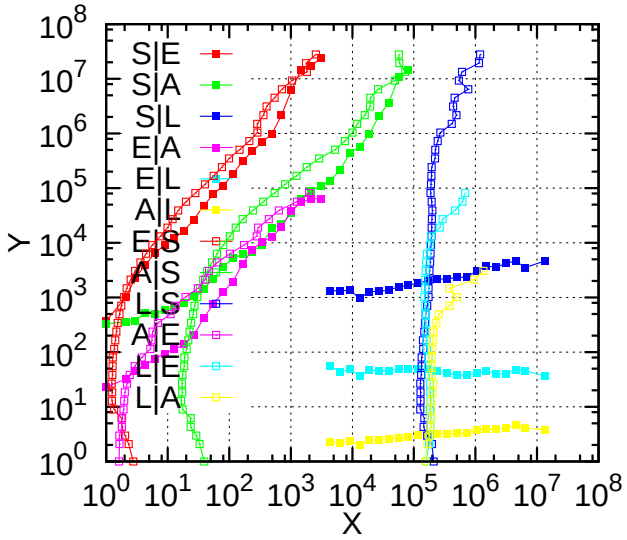

図 18: X で条件つけした Y の平均値 $\langle Y|X\rangle$ (■) と Y で条件つけした X の平均値 $\langle Y|X\rangle$ (□).

に注目する．S, A, E, L のそれぞれを

$$x_X = \frac{\log(X) - \log(X_{\min})}{\log(X_{\max}) - \log(X_{\min})} \quad X = S, A, E, L$$

のように標準化し (ただし，$X_{\max}(X_{\min})$ は X の最大 (小) 値とする)，x_A, x_E, x_L 空間において予測したい点の k 個の最近傍の標本を取りそれらの平均値 $\langle x_S\rangle$ を予測値とする (k 近傍法)．k の値を決めるために，全データを N 等分し，$N-1$ 群のみを既知とし，残り 1 群をテストに用いて予測誤差を計算した (cross validation)．N 種類の組み合わせに対して予測誤差を計算した結果，$k = 100$ あたりで最小に

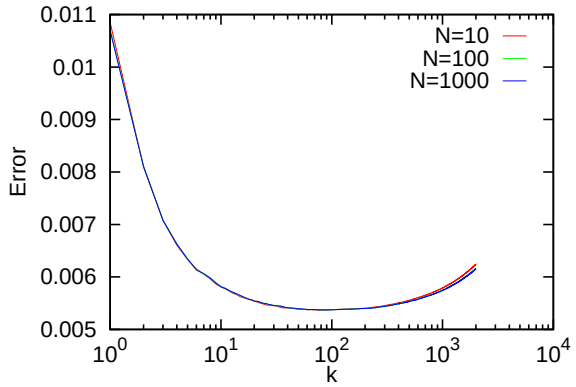

図 19: 予測誤差の k 依存性 (ただし，全店舗からランダムに 10,000 店選んで行った結果)

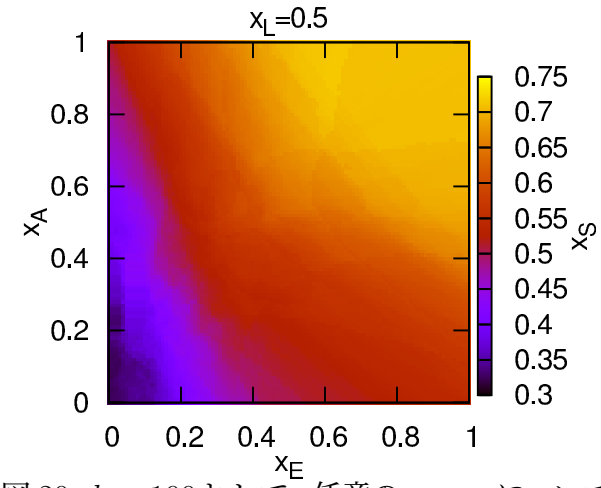

図 20: $k = 100$ として，任意の x_A, x_E について販売額 x_S を推定した結果 (ただし，$x_L = 0.5$)

なることが分かった (図 19)．このようにして，任意の店 (任意の A, E, L) について，販売額を推定する生産関数を数値的に抽出することができた (図 20)．

(2) 当初計画の達成状況について

不動産バブルの度合いを定量化する手法を確立し，アベノミクス効果の観測に役立てることができた．金融機関の振込情報を入手し，現実のお金の流れからみた企業間ネットワークの統計性を明らかにすることができた．ビジネスニュースのテキスト時系列を分析し，ニュース時系列の日中変動と地域依存性を明らかにした．売り場面積，従業員数，地価から国内の店舗の販売額を算出する生産関数を数値的に求めた．

以上の通り，様々な現実のデータについて，データを収集・入手し，分析できる形に整理し，実データに基づいて解析するという当初の計画を達成することができた．また，実証科学の視点に基づく分析により，新たな知見を得ることができた．国際会議・国内会議発表を積極的に行い，研究の方向性や位置付けを確認しながら進めることができた．

4. 今後の展望

不動産バブルについては，リアルタイムにバブル度を算出するアプリケーションの作成，海外の不動産データへの適用，バブルの発生・波及・崩壊パターンの抽出，現実を正確に表す住宅価格指数の開発が考えられる．企業間の振込ネットワークについては，企業の属性情報とネットワーク構造の関係解明，ネットワークの時間変化を考慮したテンポラルネットワークとしての解析が考えられる．これまでの企業評価は，会計や株価といった企業単体のみのデータを用いて行うことが多かったが，企業間のつながり (ネットワーク構造) もきちんと加味した上で企業を評価する手法が開発できれば，融資などの金融サービスをより向上させることができると期待される．ニュース時系列については，ニュースに対する金融市場 (為替，株，先物など) の反応に注目したより詳細な解析 (どのようなニュースが流れると価格が変動するか) が考えられる．商業統計メッシュデータについては，他の年のデータも用いた時間変化の解析，集積の利益といった外部性の分析，人口メッシュデータを用いた人の流れとの関係の分析が考えられる．

実証科学の視点から現実の様々なビッグデータを解析することがますます重要になってきている．超並列計算を活用することで，社会・経済現象のモデル化を行い，予測や制御に役

立てることを目指した研究を進めていきたい.

5. 研究成果リスト

(1) 学術論文

- 大西立顕, 石井晃, "企業間振込ネットワークにおけるページランクと振込総額の統計性", 統計数理研究所共同研究リポート, 印刷中
- 大西立顕, "経済ネットワークの数理", 横幹, vol.7, no.2, pp.100–107, 2013
- Eduardo Viegas, Misako Takayasu, Wataru Miura, Koutarou Tamura, Takaaki Ohnishi, Hideki Takayasu, Henrik Jeldtoft Jensen, "Ecosystems perspective on financial networks: Diagnostic tools", Complexity, vol.19, no.1, pp.22–36, 2013
- Ryohei Hisano, Didier Sornette, Takayuki Mizuno, Takaaki Ohnishi, Tsutomu Watanabe, "High quality topic extraction from business news explains abnormal financial market volatility", PloS one, vol.8, no.6, e64846, 2013
- Takaaki Ohnishi, Takayuki Mizuno, Chihiro Shimizu, Tsutomu Watanabe, "Detecting Real Estate Bubbles: A New Approach Based on the Cross-Sectional Dispersion of Property Prices", CARF Working Paper, CARF-F-313, 2013

(2) 国際会議プロシーディングス

該当なし

(3) 国際会議発表

- Takaaki Ohnishi, Takayuki Mizuno, Chihiro Shimizu, Hiroshi Iyetomi, Tsutomu Watanabe, "Use of house price distribution for estimating a local bubble indicator", 11th German Probability and Statistics Days 2014, Ulm, Germany, 2014 年 3 月
- Takaaki Ohnishi, Akira Ishii, "Network analysis of inter-firm payment flows", ESHIA Winter Workshop 2013, Singapore, 2013 年 11 月
- Takaaki Ohnishi, Takayuki Mizuno, Chihiro Shimizu, Hiroshi Iyetomi, Tsutomu Watanabe, "Measuring the likelihood of housing bubbles: a spatio-temporal analysis", European Conference on Complex Systems 2013, Barcelona, Spain, 2013 年 9 月
- Takaaki Ohnishi, Takayuki Mizuno, Chihiro Shimizu, Hiroshi Iyetomi, Tsutomu Watanabe, "The Mechanism behind Power-Law Distribution of Housing Prices", Statphys 25, Seoul, Korea, 2013 年 7 月

(4) 国内会議発表

- 大西立顕, 石井晃, "企業間の振込ネットワークの統計性", 日本物理学会第 69 回年次大会, 東海大学, 2014 年 3 月
- 大西立顕, 石井晃, 戸谷圭子, "金融機関振込情報のネットワーク分析", 平成 25 年度統数研研究会「経済物理学とその周辺」第 2 回研究会, 統計数理研究所, 2014 年 3 月
- 大西立顕, 水野貴之, 清水千弘, 渡辺努, "不動産市場におけるバブルの時空間構造の定量化", 第 5 回横幹連合総合シンポジウム, 香川大学, 2013 年 12 月
- 大西立顕, 水野貴之, 清水千弘, 家富洋, 渡辺努, "住宅価格のベキ分布とバブルの定量化", 日本物理学会 2013 年度秋季大会, 徳島大学, 2013 年 9 月
- 大西立顕, 水野貴之, 清水千弘, 家富洋, 渡辺努, "東京圏における住宅バブルの定量化", 平成 25 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, キヤノングローバル戦略研究所, 2013 年 9 月

(5) その他（特許, プレス発表, 著書等）

該当なし